Mixing Statistical and Symbolic Approaches for Chemical Names Recognition

Florian Boudin[♮], Juan Manuel Torres-Moreno[♮,♭] and Marc El-Bèze[♮]

[♮]Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, BP1228
84911 Avignon Cedex 9, France
[♭] École Polytechnique de Montréal - Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7
Montréal (Québec), Canada.
{florian.boudin,juan-manuel.torres,marc.elbeze}@univ-avignon.fr
http://www.lia.univ-avignon.fr

Abstract. This paper investigates the problem of automatic chemical Term Recognition (TR) and proposes to tackle the problem by fusing Symbolic and statistical techniques. Unlike other solutions described in the literature, which only use complex and costly human made ruled-based matching algorithms, we show that the combination of a seven rules matching algorithm and a naïve Bayes classifier achieves high performances. Through experiments performed on different kind of available Organic Chemistry texts, we show that our hybrid approach is also consistent across different data sets.

Key words: Term Recognition, Text Mining, Chemical Informatics.

1 Introduction

Over one million new chemical compounds are discovered and published annually. As in many scientific domains, the Organic Chemistry (OC) data are not published coherently but scattered through thousands of different journal articles. Identifiying and extracting chemical compounds is a critical task for chemical information retrieval. Information extraction technology arose in response to the need for efficient processing of documents in specialized domains. Classical Natural Language Processing (NLP) tools such as parsers, taggers or chunkers achieve very poor on OC documents. This is due to the specificity of the domain, a very wide vocabulary, long sentences containing a high quantity of ”hapax legomen”[1]. Scientists, especially chemists, want to be able to search for articles related to particular chemical compounds. Nowadays, search engines mainly depend on the "classical" title, author(s) and keywords scheme searching. Extracting chemicals from texts and using them to classify, organize and accelerate the information access fit to a wide range of possible applications.

[1] Terms which only appears once in a text.